用例に基づく形態素解析の検討*

4W－11

村上仁一†
NTT 情報通信研究所‡
〒239 神奈川県横須賀市光の丘 1-1 §

1　はじめに

形態素解析は、従来から対話、翻訳、校正などの目的のために、自然言語処理研究の一つの分野として研究が続けられている。しかし、実際の日本語では単語の境界が明確でないことや単語の多品詞性のため、多くの問題点が残っている。特に名詞連続複合語には問題が多い。本稿では従来の単語辞書の代わりに既に形態素解析されたデータを利用する、用例に基づく形態素解析を提案する。形態素解析済みのデータを利用することにより、形態素解析の精度が向上することが予想される。実際に企業名の形態素解析の実験を行なった。この結果について述べる。

2　用例に基づく形態素解析方法

本論文では、企業の検索を目的とした形態素解析を対象にした。企業を検索する際、予め形態素解析で分かち書きをすれば、無駄な検索が削減できる。特に PB 自動電話番号案内 [1] のように、あ行が 1 のボタンに縮退させて検索する方法では、大幅に候補数が減少する。

図 1に、通常の形態素解析のアルゴリズムを示す。

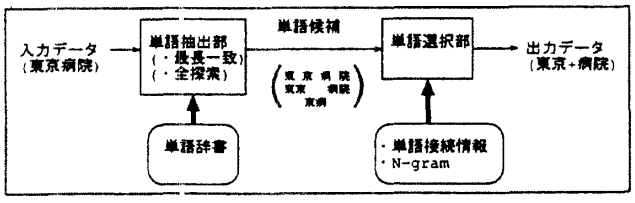

図 1: 形態素解析のフローチャート

本稿で提案する用例に基づく形態素解析は、図 1 の単語辞書の代わりに形態素解析済みのデータを利用する方法である。

単語辞書の例を表 1にあげる。

表 1: 単語辞書

1	日本+電信+電話+株式+会社
2	東京+海上+火災
3	安田+海上+火災
4	総務 | 部
5	秘書 | 部

表中、"+" は単語境界、"|" は接尾境界を意味する。

3　実験条件

1. 実験データ

実験データの作成方法を以下に示す。

(a) 電話帳から約 440 万件の企業のデータベースを作成。

(b) 作成した全データ (約 440 万件) に対し、人間が単語境界および接尾境界を付与。

(c) ランダムに 1 万件を抽出。test データとして使用。

(d) 残りのデータ (約 439 万件) を、単語辞書および trigram の連鎖確率値の計算に使用。

2. 単語抽出部

単語抽出部のアルゴリズムとして、最長一致法や文節数最小法などが良く知られている [2]。この他に全ての単語候補を検索する全探索法がある。本論文では、最長一致法と全探索法で実験を行なった。

3. 単語選択部

単語抽出部では、大量の候補が出力されるため、通常、言語情報を持ちいてこれらの曖昧さを削除している。本論文では、漢字仮名の trigram [3] を用いて候補を選出した。なお、単語境界や接尾境界は 1 文字とみなして連鎖確率を計算した。

---

*Example Based Morphological Analysis
†Jin'ichi Murakami
‡NTT Information and Communication System Lab.
§1-1 Hikarinooka Yokosuka-shi Kanagawa 239 Japan

4 実験結果

実験では、人間による形態素解析結果と比較して一致するものを正解とした。1万件の実験結果を表2に示す。

表 2: 実験結果

	最長一致	全探索
解析可能件数	9380 件	9483 件
累積正解率 1	4942 件	4785 件
～ 2	5119 件	5802 件
～ 4	5119 件	5948 件
～ 8	5119 件	5964 件
正解候補なし	4250 件	2967 件

表2中、解析可能件数とは、単語抽出部において単語候補を抽出できた件数である。

この実験結果から、第1位の正解率が最長一致で49% 全探索で47%の形態素解析精度が得られ、第4位までの累積正解率で各々51% 59%の形態素解析精度が得られることが示された。また、単語抽出部において形態素解析ができない企業名が多いこともわかった。この原因を解析した結果、単語辞書に「店」「商店」などの接尾語がないことが原因である場合が多かった。そこで、接尾語を単語辞書に登録して実験を行なった。この結果を表3に示す。

表 3: 形態素解析済み企業名+接尾語

	最長一致	全探索
解析可能件数	9711 件	9751 件
累積正解率 1	5375 件	4908 件
～ 2	5619 件	5844 件
～ 4	5625 件	5953 件
～ 8	5625 件	5966 件
正解候補なし	4003 件	3344 件

上記の実験から、単語抽出部において出力される単語候補に正解候補がないことが多いことが示された。そこで、形態素解析された企業名を、単語に分割し、全ての単語を単語辞書に登録して実験を行なった。(例えば、東京+海上+火災の場合、東京, 海上, 火災の3単語に分割して登録) この結果を表4に示す。

表 4: 形態素解析済み企業名+接尾語+単語

	最長一致	全探索
解析可能件数	9857 件	9875 件
累積正解率 1	6677 件	6119 件
～ 2	8017 件	7734 件
～ 4	8066 件	7990 件
～ 8	8067 件	8035 件
正解候補なし	1621 件	903 件

この実験結果から、第1位の正解率が最長一致で67% 全探索で61%の形態素解析精度が得られ、第4位までの累積正解率で各々81% 80%の形態素解析精度が得られることが示された。

5 考察

今回の実験から、単純な用例ベースの形態素解析では、単語辞書に登録される件数が少ないため、解析できない場合が多く出現する。そのため、解析精度が低くなり、用例ベース形態素解析の利点が得られないことが示された。しかし、今後、単語辞書に登録する形態素解析済みデータが多くなると、改善される可能性があると考えられる。

また、単語抽出部において、最長一致法のほうが、全探索法より正解率が高いことが示された。これは単語候補の選択に trigram を利用したためと考えられ、今後、4-gram などのより高度な情報を使用することにより、全探索法のほうが高くなると考えている。

6 まとめ

本論文では、既に形態素解析されたデータを用いて形態素解析をする、用例ベース形態素解析方法について述べた。そして形態素解析の実験を行ない、精度を報告した。

参考文献

[1] M. Higasida, "A Fully Automated Directory Assistance Service that Accommodates Degenerated Keyword Input via Telephones", Pacific Telecommunications Conference pp,175-179 (Jan. 1997).

[2] 長尾 真, "日本語情報処理," 社団法人電子通信学会, pp.63-64 (1984).

[3] 村上 仁一, "漢字かなの trigram をもちいたかな漢字変換方法", 情処第 43 回全大, 7H-3, pp. 3.287-288, (1991-10).